Estimation of Normal Mixtures in a Nested Error Model With an Application to Small Area Estimation of Welfare

Roy van der Weide (jointly with Chris Elbers)

DECPI - Poverty and Inequality Research Group

The World Bank

rvanderweide@worldbank.org

SAE Conference 2013, Bangkok, September 2

Outline

- Small area estimation of poverty
- *Non-Normal Non-EB* versus *Normal EB* estimation
- This study: *Non-Normal EB* estimation
  - Mixture-distributions for nested errors
  - Implications for EB estimation
- Simulation experiment
- Empirical example: Minas Gerais, Brazil, in 2000
- Concluding remarks

A measure of income poverty

- Let y_{ah} denote log income (or consumption) for household h residing in area a, and let s_{ah} denote the household size.
- Let y_a and s_a be vectors with elements y_{ah} and s_{ah}, respectively.
- The objective is to determine the level of welfare for small area a which can be expressed as a function of y_a and s_a: $W(y_a, s_a)$.
- The welfare function is typically non-linear.
- A popular example is the share of individuals whose income falls below the poverty line:

$$W = \frac{1}{N_a} \sum_h s_{ah} 1(y_{ah} < Z), \tag{1}$$

where N_a denotes the number of individuals in area a.

Estimating poverty

- Suppose that household level (log) income can be described by:

$$y_{ah} = x_{ah}^T \beta + u_a + \varepsilon_{ah} \tag{2}$$

- Suppose that we have data on x_{ah} for all households (from the population census), but observe y_{ah} only for a small subset of the population (from an income survey).

- Consider $\hat{\mu}_a$ as an estimator for $W(y_a, s_a)$:

$$\hat{\mu}_a = \frac{1}{R} \sum_{r=1}^{R} W\left(\tilde{y}_a^{(r)}, s_a\right), \tag{3}$$

  where $\tilde{y}_{ah}^{(r)} = x_{ah}^T \tilde{\beta}^{(r)} + \tilde{u}_a^{(r)} + \tilde{\varepsilon}_{ah}^{(r)}$.

ELL (2003) versus Molina and Rao (2010)

- **Elbers, Lanjouw and Lanjouw (2003, Econometrica):**
  - More flexible: Permits non-normal errors
  - Estimates the distributions for u_a and ε_{ah} non-parametrically
  - But does not take full advantage of all available data (do not adopt EB estimation)
- **Molina and Rao (2010, Canadian Journal of Statistics):**
  - Does adopt EB estimation
  - But is less flexible: Assumes normal errors

The distribution matters when estimating poverty

- Getting the error distributions right is not merely a matter of efficiency.
- Getting the distributions wrong will introduce a bias.
- Whether the magnitude of this bias is meaningful in practice is an empirical question.
- Choice between *non-normal non-EB* and *normal-EB* is motivated by:
  - The degree of non-normality found in the data.
  - How much information one stands to ignore by not adopting EB.
- The latter is largely determined by:
  - The number of areas that are covered by the survey.
  - The size of the area random effect.

The objectives of this study

- The approach developed in this study aims to combine the best of both worlds.
- We adopt EB estimation.
- Without restricting the distributions of the errors.

Normal mixtures in a nested error model

- Let the probability distribution functions for u_a and ε_{ah} be denoted by F_u and G_ε.

- Consider normal-mixture distributions as a flexible representation of F_u and G_ε:

$$F_u = \sum_{i=1}^{i=m_u} \pi_i F_i \quad (4)$$

$$G_\varepsilon = \sum_{j=1}^{j=m_\varepsilon} \lambda_j G_j. \quad (5)$$

- We assume that F_i and G_j are normal distribution functions with means μ_i and ν_j, and variances σ_i^2 and ω_j^2.

Estimation of normal-mixtures in a nested error model

- Let $e_{ah} = y_{ah} - x_{ah}^T \beta$, and $\bar{e}_a = \bar{y}_a - \bar{x}_a^T \beta$.
- We have:

$$e_{ah} = u_a + \varepsilon_{ah} \tag{6}$$

$$\bar{e}_a = u_a + \bar{\varepsilon}_a. \tag{7}$$

- The challenge here lies in the nested error structure: We wish to estimate the distribution functions for u_a and ε_{ah}, but we observe neither directly.
- For details on our method of estimation, please see the presentation by Chris Elbers tomorrow.

EB with normal mixture distributions

- It follows that $p(u_a|\bar{e}_a)$ is a normal mixture with known parameters whenever $p(u_a)$ and $p(\varepsilon_{ah})$ are normal mixtures.

- The conditional mean solves:

$$E[u_a|\bar{e}_a] = \sum_i \alpha(\bar{e}_a)\left(\gamma_{ai}\bar{e}_a + (1-\gamma_{ai})\mu_i\right), \tag{8}$$

  where $\gamma_{ai} = \sigma_i^2/(\sigma_i^2 + \sigma_\varepsilon^2/n_a)$, and where $\alpha(\bar{e}_a)$ denote the mixing probabilities of $p(u_a|\bar{e}_a)$.

- Note that normal-EB is nested as a special case, where:

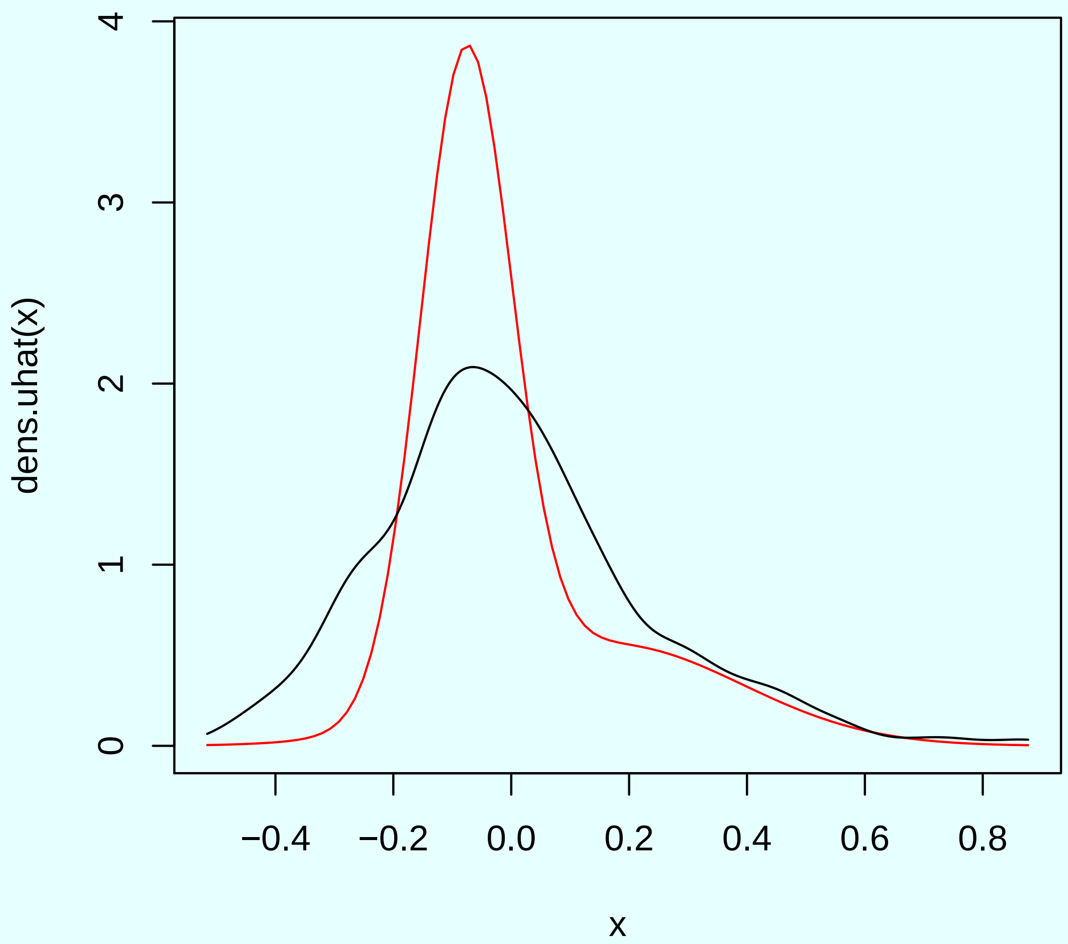
$$\begin{aligned} E[u_a|\bar{e}_a] &= \gamma_a \bar{e}_a \\ var[u_a|\bar{e}_a] &= (1-\gamma_a)\sigma_u^2, \end{aligned}$$

  with $\gamma_a = \sigma_u^2/(\sigma_u^2 + \sigma_\varepsilon^2/n_a)$.

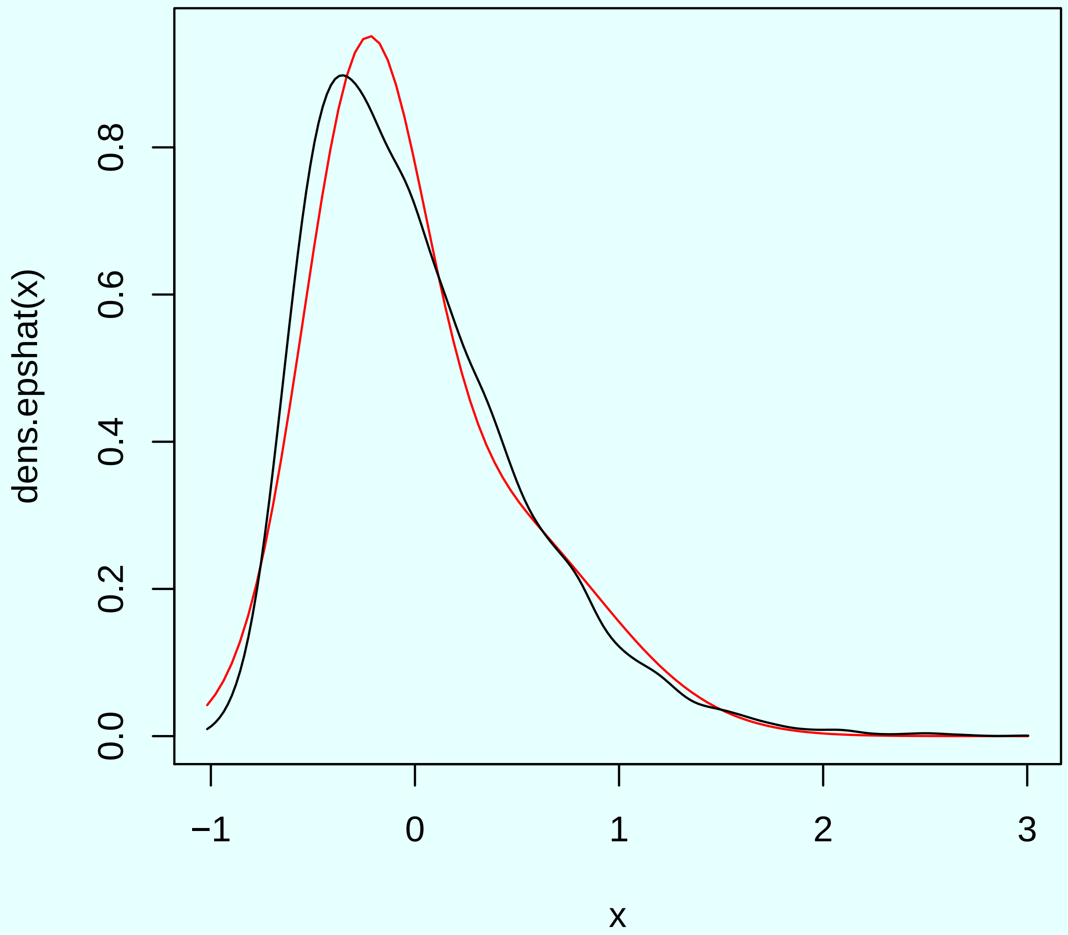
A small simulation experiment

- We simulate a census population with 500 areas, and $15 * 200 = 3000$ households in each area.
- The survey samples 15 households from each of the 500 areas.
- $\sigma_e^2 = 0.3$, and $\sigma_u^2/\sigma_e^2 = 0.1$, which yields: $\sigma_u^2 = 0.03$ and $\sigma_\varepsilon^2 = 0.27$.
- $u_a \sim skew{-}t(0, scale = 1, skew = 3, df = 6)$, and $\varepsilon_{ah} \sim skew{-}t(0, scale = 1, skew = 6, df = 24)$. (Both u_a and ε_{ah} are standerdized so that they have mean 0 and variances 0.03 and 0.27, respectively.)
- There is one regressor, x_{ah} with $\mu_x = 0$ and $\beta = 1$. We set $R^2 = 0.4$, so that $\sigma_x^2 = R^2\sigma_e^2/(\beta^2(1 - R^2)) = 0.2$.
- Overall poverty is estimated at 32.6 percent.

A small simulation: Estimating F_u



A small simulation: Estimating G_ε



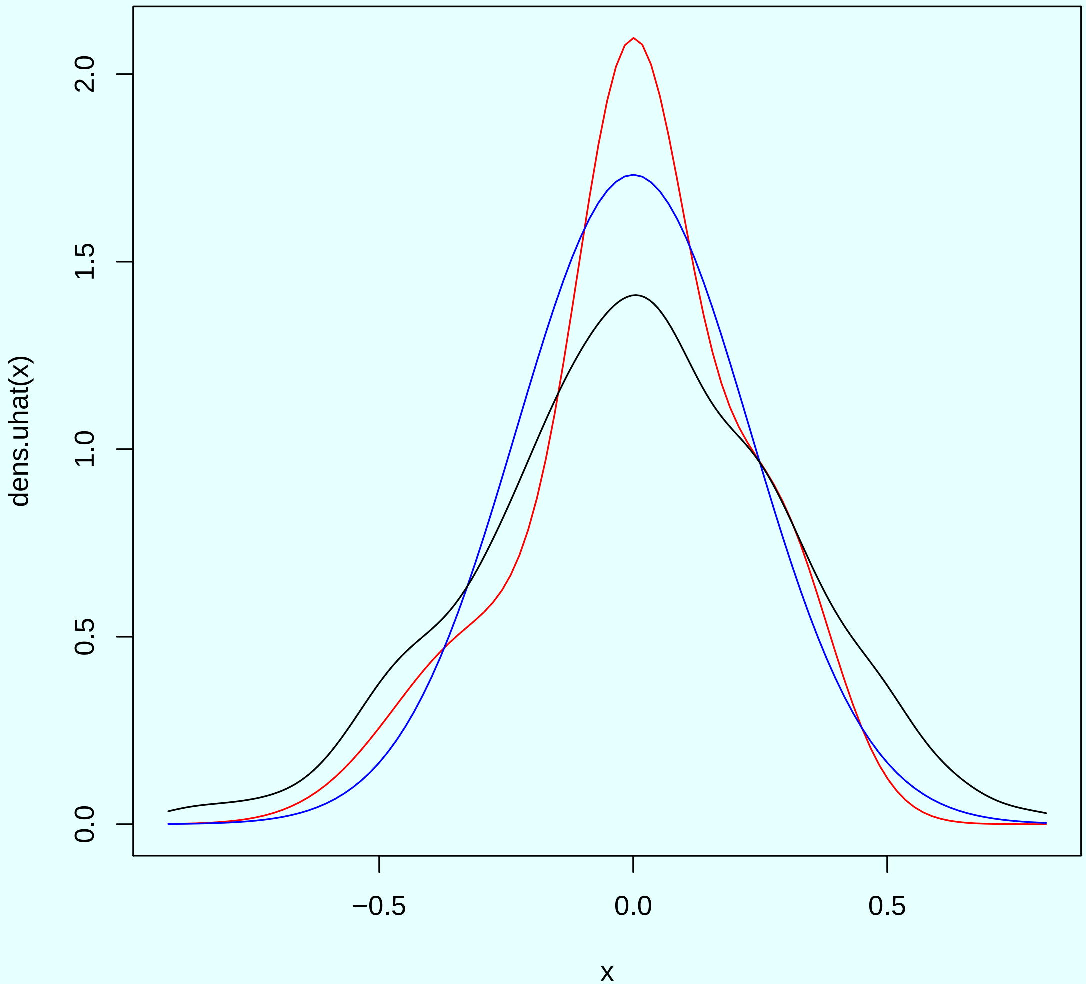
A small simulation: Bias and RMSE

- Non-EB:
  - Bias: $-\mathbf{1.61}$ (N) versus $-\mathbf{0.20}$ (NM).
  - RMSE: $\mathbf{9.27}$ (N) versus $\mathbf{9.13}$ (NM).
- EB:
  - Bias: $-\mathbf{0.94}$ (N) versus $\mathbf{0.30}$ (NM).
  - RMSE: $\mathbf{5.66}$ (N) versus $\mathbf{5.38}$ (NM).
- Normal mixture does better than normal errors, but the improvement is modest.

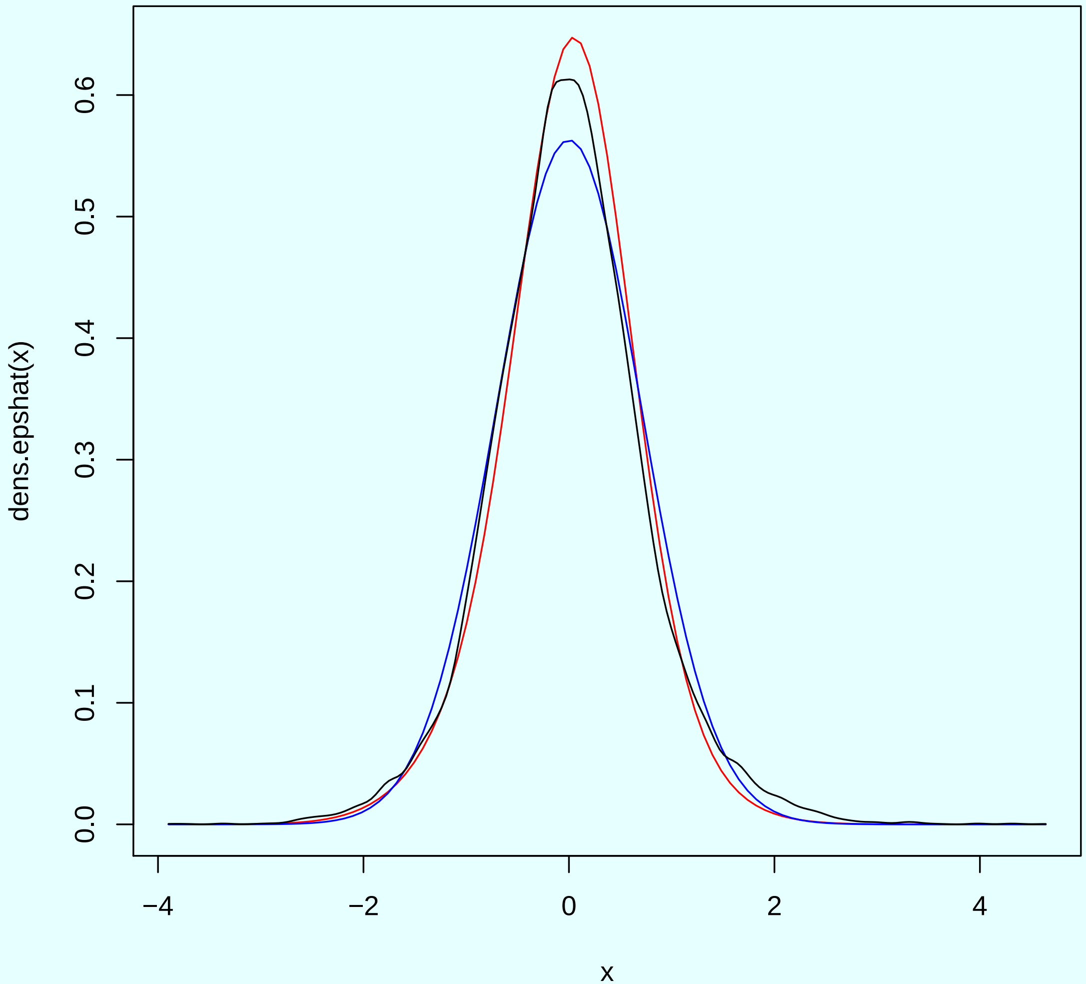
An application to Brazil: Bias and RMSE

- We use 12.5% of the 2000 population census of Minas Gerais, Brazil, which amounts to approx. $600,000$ households divided over 853 municipalities.
- An artificial survey is obtained by sampling 15 households from each of the 853 municipalities.
- The regression model consists of 12 independent variables on demographics and education, which yields an adjusted-R^2 of 0.423.
- The location effect is estimated at: $\hat{\sigma}_u^2/\hat{\sigma}_e^2 = 0.097$.
- The overall poverty rate is estimated at 22.2 percent.

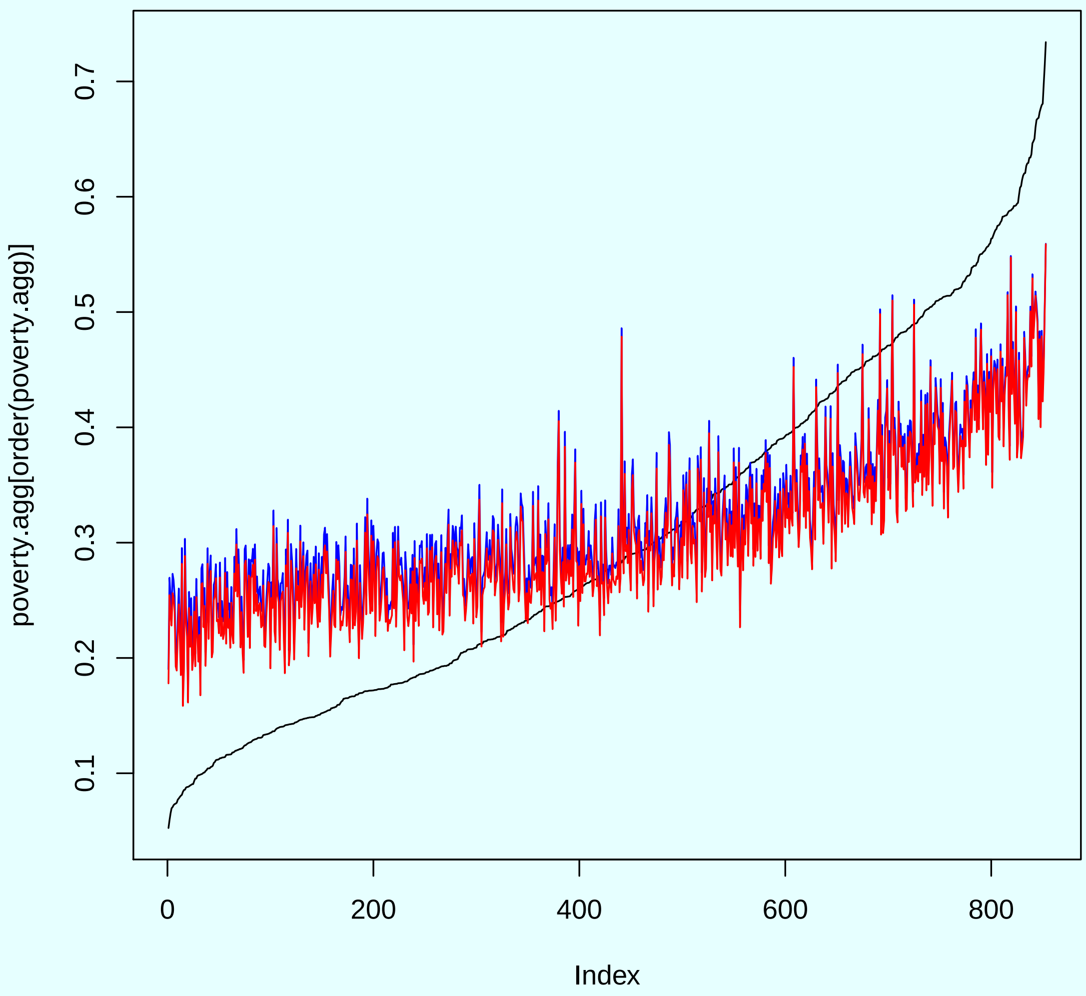
An application to Brazil: F_u



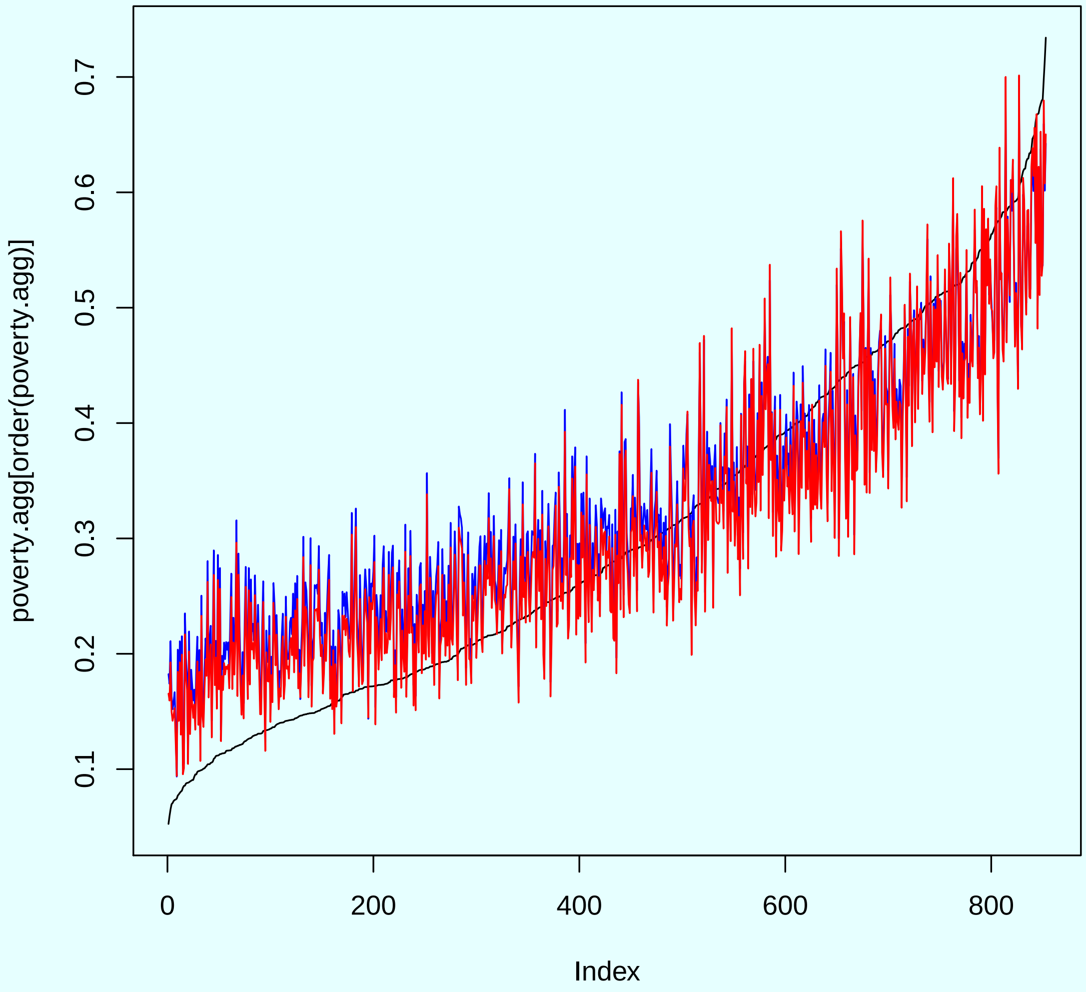
An application to Brazil: G_ε



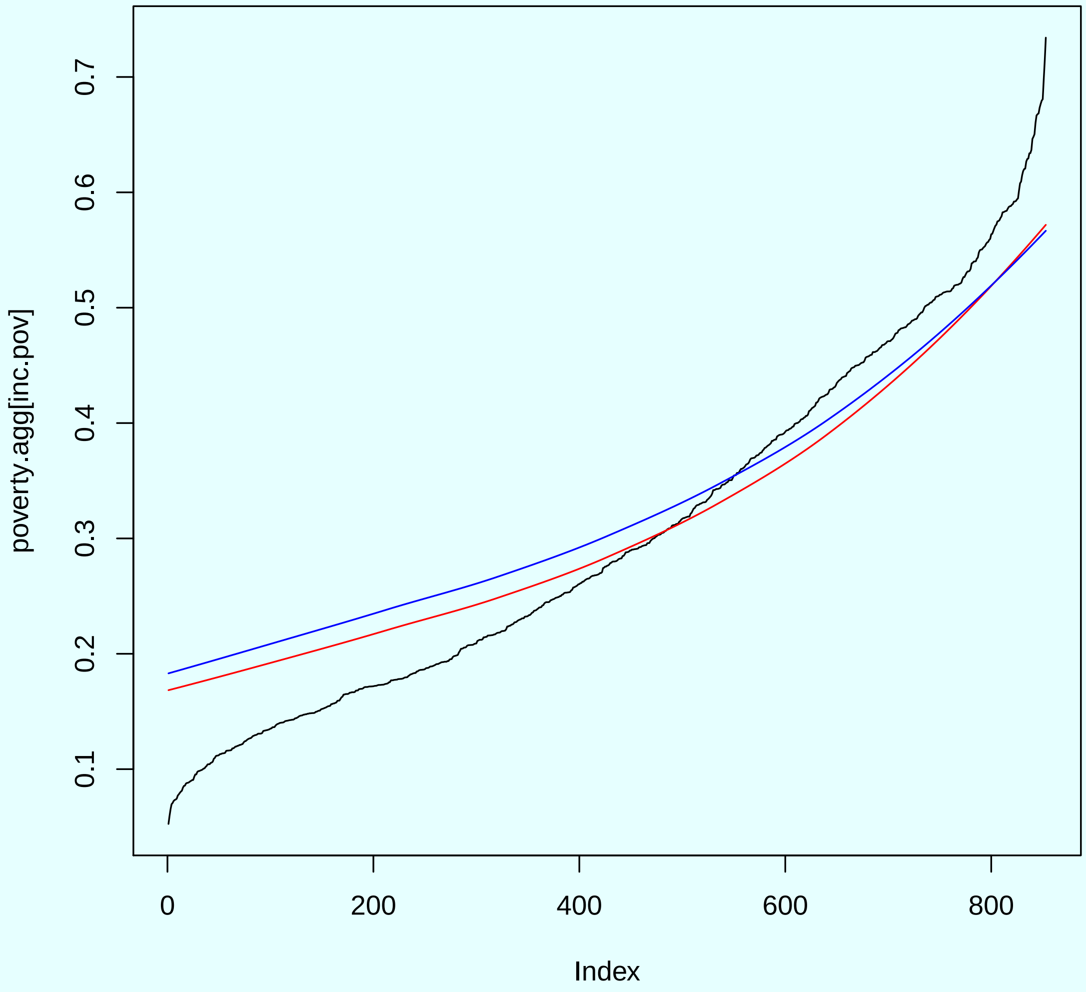
An application to Brazil: non-EB estimates



An application to Brazil: EB estimates I



An application to Brazil: EB estimates II



An application to Brazil: Bias and RMSE

- Non-EB:
  - Bias: 1.37 (N) versus 0.10 (NM).
  - RMSE: 10.06 (N) versus 9.84 (NM).
- EB:
  - Bias: 2.17 (N) versus 0.78 (NM).
  - RMSE: 7.00 (N) versus 6.62 (NM).